Developing Academic English Language Proficiency Prototypes for 5th Grade Reading: Psychometric and Linguistic Profiles of Tasks An Extended Executive Summary

CSE Report 720

Alison L. Bailey, Becky H. Huang, Hye Won Shin, and Tim Farnsworth

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

Frances A. Butler

Language Testing Consultant

June 2007

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

DEVELOPING ACADEMIC ENGLISH LANGUAGE PROFICIENCY PROTOTYPES FOR 5TH GRADE READING: PSYCHOMETRIC AND LINGUISTIC PROFILES OF TASKS AN EXTENDED EXECUTIVE SUMMARY[1]

Alison L. Bailey, Becky H. Huang, Hye Won Shin, and Tim Farnsworth
National Center for Research on Evaluation, Standards, and Student Testing
University of California, Los Angeles

Frances A. Butler
Language Testing Consultant

Abstract

Within an evidentiary framework for operationally defining academic English language proficiency (AELP), linguistic analyses of standards, classroom discourse, and textbooks have led to specifications for assessment of AELP. The test development process described here is novel due to the emphasis on using linguistic profiles to inform the creation of test specifications and guide the writing of draft tasks. In this report, we outline the test development process we have adopted and provide the results of studies designed to turn the drafted tasks into illustrative prototypes (i.e., tried out tasks) of AELP for the 5th grade. The tasks use the reading modality; however, they were drafted to measure the academic language construct and not reading comprehension per se. That is, the tasks isolate specific language features (e.g., vocabulary, grammar, language functions) occurring in different content areas (e.g., mathematics, science, and social studies texts). Taken together these features are necessary for reading comprehension in the content areas. Indeed, students will need to control all these features in order to comprehend information presented in their textbooks. By focusing on the individual language features, rather than the subject matter or overall meaning of a text, the AELP tasks are designed to help determine whether a student has sufficient antecedent knowledge of English language features to be able to comprehend the content of a text.

The work reported here is the third and final stage of an iterative test development process. In previous National Center for Research on Evaluation, Standards, and Student

---

[1] We gratefully acknowledge the following publishers for permission to use textbook excerpts in the CRESST test development process: Harcourt for *Math* (2002) National Edition, *Science* (2000) California Edition and *Social Studies: Early United States* (2002) National Edition; Houghton Mifflin for *Mathematic* (2002) California Edition*; Science* (2000) California Edition, *Social Studies: America Will Be* (1999) National Edition; McGraw-Hill for *Math Explorations and Applications* (2003) National Edition, *Science* (2000) California Edition, *United States: Adventure in Time and Place* (2001) National Edition.

Testing (CRESST) work, we conducted a series of studies to develop specifications and create tasks of AELP. Specifically, we first specified the construct by synthesizing evidence from linguistic analyses of ELD and content standards, textbooks (mathematics, science, and social studies), and teacher talk in classrooms, resulting in language demand profiles for the 5th grade. After determining task format by frequency of assessment types in textbooks, we then created draft tasks aligned with the language profiles.

The goals of the current effort were to take these previously drafted tasks and create prototypes by trying out the tasks for the first time with 224 students from native English and English language learner (ELL) backgrounds. Students across the 4th-6th grades, as well as native-English students, are included in the studies because native speakers and adjacent grades provide critical information about the targeted language abilities of mainstream students at the 5th grade level. Phase 1 (n= 96) involved various tryouts of 101 draft tasks to estimate duration of administration, clarity of directions, whole-class administration procedures, and an opportunity to administer verbal protocols to provide further information about task accessibility and characteristics. Phase 2, the pilot stage, involved administration of 40 retained tasks (35 of which were modified as a result of Phase 1) to students in whole-class settings (n=128). Analyses included item difficulty and item discrimination. The rationale for retaining or rejecting tasks is presented along with psychometric/linguistic profiles documenting the evolution of example effective and ineffective prototype tasks. The final chapter of the report reflects on the lessons learned from the test development process we adopted and makes suggestions for further advances in this area.

Overview and Outline of the Report

The work described in the full report is the culmination of several years of research at the national Center for Research in Evaluation, Standards, and Student Testing (CRESST) that focused initially on articulation of the academic English construct in school settings, and finally on the use of that information for the development of prototype reading tasks of academic English. Specifically, the report presents findings from a series of small-scale try-outs and a pilot study with reading tasks designed to assess 5th grade academic English language proficiency (AELP).

The report begins with a summary of the prior research at CRESST which provides the background and context for the AELP task development. The specific goals of the task development effort are then outlined. Next, we describe the procedures and instrumentation of each of the two phases of administering and revising the AELP tasks, followed by analyses of the data collected during in the pre-pilot phase and the subsequent pilot phase.

Six tasks profiles demonstrate how tasks were refined in light of feedback from verbal protocols with students and psychometric information on item-level performance. Tasks based on reading passages from mathematics, science, and social studies content areas are used to illustrate in considerable depth the decision-making process for how tasks could be retained without modification, modified and retained for piloting, or rejected as unsuitable for further development. The report concludes with recommendations for refinement of the research and standards-informed test development process and implications for further research in this area.

Context and Stages of AELP Test Development

The impetus for this long-term initiative grew out of the need to ensure access for all students in evaluation of their academic progress. In the mid to late 1990's, the validity of large-scale (standardized) assessments with English language learner (ELL) students came into question (August & Hakuta, 1997; Butler & Stevens, 1997, 2001; LaCelle-Peterson & Rivers, 1994). This concern led to further issues, including the use of test accommodations with ELL students (Abedi, 1997; Abedi, Lord, & Plummer, 1997; Butler & Stevens, 1997) and the effectiveness of existing language proficiency tests for evaluating the English language skills of those students (Stevens, Butler, & Castellon-Wellington, 2000; Butler & Stevens, 2001; Bailey & Butler, 2002/2003). CRESST research was showing that existing language tests were not good predictors of performance on standardized content tests (Butler & Castellon-Wellington, 2000/2005). There was a mismatch between the language tested on language proficiency tests (every-day vocabulary and simple structures) and the language used on content tests and in the classroom (more precise uses of vocabulary and complex structures; Stevens, Butler, & Castellon-Wellington, 2000). The distinctions between the two are typically characterized as social versus academic English, although the distinctions are not always easy to articulate. Since both are critical to the student's English language development, educators began to recognize the need for expanding the content domain of K-12 English language proficiency tests to include academic English.

The *No Child Left Behind Act of 2001*, which required that ELL students show measurable yearly progress in English language development (ELD), brought the language proficiency assessment of ELL students to the forefront of the national educational discussion. The need for language tests that focused on academic English, or at least included features of academic English in the test content, rapidly

became apparent because of the high stakes decisions (e.g., redesignation of ELL students) being made on the basis of student performance on language tests, and the accountability of schools and states for student performance. Many existing commercial ELD tests failed to capture the language demands required for academic success, thus motivating the development of AELP assessment tasks.

Our test development approach follows the National Research Council's call for evidence-based educational research as well as established procedures in test development, such as conducting a needs analysis, to ensure high technical quality (e.g., American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999; Bachman, 1990; Davidson & Lynch, 2002)[2]. Consequently, in previous CRESST work we conducted a series of studies to develop specifications and create draft tasks of AELP. Specifically, we specified the construct by synthesizing evidence from linguistic analyses of ELD and content standards, textbooks (mathematics, science, and social studies), and teacher talk in classrooms, resulting in language demand profiles for the 5th grade (Bailey, Butler, LaFramenta, & Ong, 2001/2004; Bailey & Butler, 2002/2003; 2006; Butler, Lord, Stevens, Borrego, & Bailey, 2003/2004; Butler, Bailey, Stevens, Huang, & Lord, 2004).

The table below is taken from Butler et al. (2004) and shows how texts from the different content areas contain different language features. The profiles were used as part of the test specifications for AELP tasks and guided the creation of draft tasks that use actual texts from mathematics, science and social studies textbooks (see Bailey, Stevens, Butler, Huang, & Miyoshi, 2005). The information in the profiles guided the prevalence of linguistic features in the tasks and also the linguistic characteristics of the text selections to which the tasks are attached.

A key contribution to test development from these efforts is the greater specificity given to the construct "academic English." Academic English language (AEL) has become one of the popular foci in the current field of education and assessment (United States Government Accountability Office [GAO], 2006). At its simplest, AEL refers to the language used for the purpose of "acquiring new knowledge and skills...imparting new information, describing abstract ideas, and

---

[1] The techniques for needs analysis grew out of work in the area of syllabus design. See McNamara (1996, p. 36) for an historical perspective on the use of needs analyses in language test development. See also Witkins and Altschuld (1995) for needs assessment techniques.

developing students' conceptual understanding" (Chamot & O'Malley, 1994, p. 40). AEL is distinct from the social language used in school (e.g., Scarcella, 2003; Schleppegrell, 2001); it encompasses the vocabulary, syntactic structures and discourse features that are "necessary for a student to access and engage with their grade-level curriculum" (Bailey & Heritage, forthcoming).

Table 1

Content Framework for Developing an Assessment of Academic Language Proficiency

Content category	Mathematics	Science	Social studies
Vocabulary			
Clause connectors	√	√	√
Non-academic vocabulary			
Academic vocabulary (AV)			
General AV (high-frequency)	√	√	√
Specialized AV (defined in context)	--	√	√
Measurement words	√	√	--
Proper nouns	--	--	√
Grammar			
Nominalizations	--	√	√
Noun phrases	√	√	√
Participial modifiers	--	√	√
Passive forms	--	√	√
Prepositional phrases	√	√	√
Organization of Text			
Comparison	√	√	√
Definition	--	√	√
Description	√	√	√
Enumeration	√	√	√
Exemplification	--	√	√
Explanation	--	√	√
Labeling	--	√	√
Paraphrase	√	√	√
Scenario	√	--	--
Sequencing	√	√	√

Note. From *Academic English in Fifth-grade Mathematics Science, and Social Studies Textbooks* (p. 110), by F.A. Butler, A.L. Bailey, R. Stevens, B. Huang, and C. Lord, 2004, CSE report # 642. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Copyright 2004 by CRESST. Reprinted with permission.

The tasks are designed to measure student knowledge of AEL through reading. They are intended to measure the academic language construct and not reading

comprehension; that is, the tasks isolate specific language features (e.g., vocabulary, grammar, language functions) of the different content areas (e.g., mathematics, science, and social studies). Taken together these features are necessary for reading comprehension in the subject areas; indeed, students will need to control all these features in order to comprehend information presented in their textbooks. By focusing on the individual language features, rather than the subject matter or overall meaning of a text, the AELP tasks are designed to help determine whether a student has sufficient antecedent knowledge of English (i.e., linguistic features such the nominalization of verbs and the complex embedding of clauses within sentences) to be able to comprehend the content of a text.

The focus of the current report was divided into two phases: a) a pre-pilot phase of initial tryouts with 101 drafted AELP tasks, which provided information that led to retention, refinement, or rejection of the tasks; and b) a pilot phase with 40 retained and largely refined tasks to create prototype tasks of AELP.

Throughout the empirical testing of tasks, accountability is maintained through the documentation of the processes of task development and modification, and ongoing qualitative and quantitative analysis of test-taker performance on the tasks. This documentation, which we conduct here as an "audit trail" (Davidson, Kim, Lee, Li, & Lopez, 2006), serves as a primary source of evidence for evaluating the overall validity of the test, as well as providing a guide for future development efforts. Starting from test specifications, test developers can use the audit trail to document how tasks change or are eliminated during the test development process due to data from pre-pilot and pilot testing, expert reviews, and revision, thus providing important information for creating a validity argument.

With the addition of psychometric and linguistic information provided by the studies conducted for the current report, these tasks can serve as potential models or prototypes for others who are developing AEL proficiency tests. The full report provides several example profiles of effective and ineffective tasks, along with linguistic profiles unique to each task.

Procedures Overview

One hundred and one draft reading tasks were created from test specifications based on linguistic analysis of mathematics, science, and social studies textbooks. A focus group of nine ESL and content-area teachers rated the passages and tasks for linguistic difficulty and item type familiarity. Minor changes were made to the tasks

based on feedback from the teachers (see Bailey et al., 2005). A phase of several tryouts provided pre-pilot feedback on administration and clarity of the tasks, and was followed by a pilot phase.

Phase 1: Pre-pilot tryouts

The pre-pilot included a) initial tryout with an in-coming fifth-grader to help identify remaining formatting issues and an estimated completion time, b) administration of the tasks to 77 students in whole-class settings, and c) a verbal protocol version of the tasks administered individually to 18 additional students. The pre-pilot sample was recruited from a university elementary laboratory school situated in a large urban area of Southern California. Students across the 4th-6th grades were included in the studies because students in adjacent grades provide critical information about the targeted language abilities of students at the 5th grade. The tasks must be harder for students in the 5th grade than for students in the 6th grade, but easier than for students in the 4th grade. Also, the items must be tried out with native speakers of English to make sure they are neither exceptionally easy, nor difficult for this population of students who are assumed to have the level of academic English proficiency which the ELL population is expected to move toward. At this stage in the process native-speaker feedback is critical. Appropriately, the majority (79%) of students in the pre-pilot were native speakers of English.

Phase 2: Pilot Study

The pilot study was designed to elevate the draft tasks to formal prototype tasks with accompanying psychometric and linguistic information on each task. The pilot was conducted with the 40 retained and predominantly modified tasks (as will be discussed, 35 of the 40 were modified as a result of the Phase 1 pre-pilot studies). One hundred and twenty-eight students were administered the draft tasks in whole-class settings. Two urban elementary schools in Southern California were recruited for the pilot study. School 1 consisted of predominantly Caucasian students (66.4%). The other major ethnic groups were Hispanic (16.1%) and Asian (8.4%). There was 6.2% ELL students, and 9.3% of the student body qualified for free/reduced-price meals. In contrast, the majority of the student population in School 2 was Hispanic (84.4%). More than half of the student body was designated as ELL students (59.2%), and a high proportion of the students qualified for free/reduced price meals (81.8%). The average Academic Performance Index (API) scores for the two

schools were 887 and 661, respectively[3]. Tests scores for the California Standards Test in English Language Arts (CST-ELA) were also available for 121 of the students. For 73 students who were designated ELL students by their districts, scores on the California English Language Development Test (CELDT) were also available. In summary, the Phase 2 student sample included native-English speaking students for the reasons mentioned above, but a larger number of ELL students was included as well to ensure information about the revised tasks from the target population.

Summary of Results

The work reported here yielded prototype tasks designed to assess academic English at the 5th grade level through the reading modality. The tasks are designed to test a range of language functions and features, not reading comprehension per se and not content knowledge, although the selection of language functions and features being tested was drawn from content material to make the tasks most relevant to language used in the academic context.

Pre-pilot Results and Refinements

The pre-pilot phase consisted of group administrations with 77 predominantly English-only 4th-6th graders and verbal protocol data from an additional 18 students distributed across these grades and representative of different reading ability levels and Spanish- and English-dominant language backgrounds. Results suggested that, of the original 101 draft tasks, 40 were sufficiently effective for retention in terms of a combination of quantitative and qualitative factors, including item difficulty, item discrimination on reading ability, distinguishing between Spanish-dominant versus English-dominant home language backgrounds, free of gender biases, and free from anomalies in directions and formatting ambiguities, or at least contained formatting and wording issues that could be refined. However, 35 of these AELP tasks required modifications of some sort (e.g., rewording of directions, etc.) before they were considered acceptable by internal review for the pilot phase.

The intent of verbal protocols at the pre-pilot phase was to gain in-depth information about the draft tasks for use in making any necessary refinements at the end of Phase 1. Using verbal protocol techniques (see Cohen, 2000), students were asked to think aloud as they answered the tasks and then asked at the completion of

---

[3] The California statewide average in 2006 was 720. (Source: http://dq.cde.ca.gov/dataquest).

the tasks which texts and tasks they found easiest or most difficult, and why. All responses there audio recorded and later transcribed. The data driven analyses provided feedback on formatting issues, clarity of directions, word-level issues, item-level issues, answer strategies, and use of background information. Specifically, results suggested that the students found passage length, prior knowledge, and familiarity with vocabulary to be particularly important for comprehension of the academic English texts. This result is not surprising in that a longer passage, for example, could be more difficult for students simply because it contains more language to process. Students also identified prior knowledge as a means of comprehending the informational load of reading passages. Students vary in what they can interpret based on their prior knowledge or experience. Similarly, comprehending a passage involves familiarity with vocabulary, which is again a function of prior knowledge. Students either knew the word or did not depending on their exposure and familiarity with the language used in the text.

The following is just one example from the data that illustrates these findings:

> "George Washington [reading passage]. Because there were a lot of words in it. A lot of things like "militia," stuff like that. First I didn't know what all that was. Stuff like that, or "Hessian", "revolution."

The observed behaviors during the verbal protocol revealed important results on the process, namely, the comprehension and cognitive behavior of the students as they attempted to understand the texts. Thus, the verbal protocol in this study served as a tool utilized in conjunction with the quantitative information to improve the AELP tasks before they were submitted to pilot testing at Phase 2.

To conclude, the Phase 1 pre-pilot allowed us to identify problematic tasks. We focused on tasks which seemed not to be working well according to item difficulty or item discrimination functions. These tasks were then later subject to extra scrutiny for refinement or rejection. Although the focus of these analyses is the individual tasks, some passages had multiple problematic tasks, raising questions about the suitability of some of the selected reading passages as well.

Results of the Pilot Study

The pilot phase was conducted using group administrations of the 40 retained and largely refined tasks with 128 4^{th}-6^{th} grade English-only and ELL students. Correlations between percent correct on the AELP tasks and state standards-based assessments of English reading and ELD were very high. The correlation with the

CST ELA assessment was r (121) = .707 (p < .0001). The correlation between the AELP percent correct and total CELDT score (a measure of listening, speaking, reading, and writing) was r (73) = .643 (p < .0001). The CELDT Reading subtest, which is closest to the AELP tasks in both construct definition and content, was even higher (r (70) = .725, p < .0001).

Item difficulty was calculated for all 40 tasks in the pilot dataset (n=128). This statistic is the proportion of test takers who got an item correct. The easiest item (Q31) had 92% of the students attempting this item correctly answering it. The most difficult item (Q17-20 cluster) had just 9% of students answering it correctly. The majority of tasks had difficulty estimates in the .50-60 range, indicating that most tasks were neither exceptionally easy nor difficult for this sample.

In the item discrimination analysis, we used student performance on the CST ELA to create two groups; masters and non-masters. Eleven tasks discriminated poorly between the two groups (i.e., D < .25). Fourteen tasks discriminated moderately well (i.e., .25 ≤ D < .35; including a new cluster item of 4 non-independent forced-choice items), and the remaining 17 discriminated adequately (D ≥ .35; including a second new cluster item).

To conclude, the pilot findings show that the majority of the draft tasks were in the middle range of difficulty. Items that are on the extremes of the difficultly continuum might be avoided as exemplar tasks, although an operational test of AELP reading at the 5th grade that had the purpose of measuring student development in AELP would require tasks at both the easy and difficult levels, as well as in the middle range. Most tasks discriminated good from poor readers in the sample moderately to adequately. Those tasks with moderate discrimination are candidates for refinement and subject to further tryouts and re-piloting. While the tasks with adequate discrimination may also require further refinements, these are certainly the most promising tasks from the pilot stage. We profile three of these tasks in the appendix of this summary.

What Makes for Effective AELP Tasks?

In the report, we present both effective and ineffective tasks from three content areas: Mathematics, Science, and Social Studies. Each profile consists of the task specifications, the target features of the AELP construct, the linguistic profiles, relevant audit trail entries, the pre-pilot and pilot results, and relevant verbal protocol excerpts. The task profiles serve as documentation of the evolution of

example AELP tasks in our test development process. We conclude each profile with a recount of the information to substantiate our argument for its role as an AELP prototype task.

What made tasks effective prototypes from our perspective was the fact that while they targeted academic English in different linguistic domains (lexical, syntactic, discourse), the measurement of specific aspects of academic English was still predominant (e.g., general academic vocabulary is the focal linguistic feature measured by the task in profile III of the appendix). The selection of these tasks presents the full range of difficulty, from quite difficult (p =. 23) to relatively easy (p = .75), which would be necessary for an operational assessment of AELP to capture if the purpose was to measure progress in academic English language skills. These tasks all distinguished between students who came from Spanish-dominant home language backgrounds and those who came from English-dominant home language backgrounds, although we caution that this supplementary language variable can only be a rough proxy for proficiency. Additionally, this variable may function as an indictor of not only language differences, but cultural differences as well; and significant differences in task performance may also be a reflection of cultural biases in the tasks. On most tasks in this study, girls outperformed boys. However, for tasks to be considered effective, we required that the difference in performance by gender be slight and always statistically non-significant. Finally, the tasks all had "adequate" discrimination indices, suggesting that they distinguished between the good and poor readers who attempted them.

Recommendations

Recommendations for Further Research

We make recommendations for future research in three main areas: First, further research can be conducted with the data collected during this stage of the AELP project. For example, at the item level, further examination of the tasks can be made in terms of difficulty. Specifically, we can investigate further characteristics (e.g., CELDT score, years in the US) of students who incorrectly answered the tasks and those who correctly answered the task, as well as examine correspondences between difficulty and specific linguistic characteristics.

Second, further research on the AELP construct and how tasks designed to assess the construct is needed at other grade levels and in other modalities. The current study targeted the 5th grade and the reading modality only. However, to

respond to the needs of students across the K-12 span and to address the demands of academic language in the areas of listening, speaking, and writing, the efforts and processes we have described here will need to be repeated to take account of all grades and the additional modalities. Opportunity to continue with this line of research is possible with new CRESST projects currently underway that focus on the validity of assessments used with ELL students.

Third, prior knowledge of the content is not, or should not be, necessary for providing the correct response for a language task; however, the verbal protocol data in this study show that the students who articulated their thought processes were, in fact, strongly influenced in completing tasks by their prior knowledge. Thus the interrelationship between language and content knowledge should not be minimized (e.g., Haladyna & Downing, 2004). Immediate further investigation is warranted to help ensure that interaction between the two does not interfere with assessment of the academic language construct.

Recommendations for the Test Development Process

The goal of the work here was to describe a process for developing tasks that tap academic English and to provide examples or prototypes of such tasks that could be used as models for similar test development efforts. The strength of the test development process we followed for this AELP project was that we can improve both the tasks *and* the process itself. Specifically, we learned from implementing this process what could be improved, and the changes we recommend are captured in the Figure 1.[4] The transparency in the process we followed was achieved through the use of an audit trail during the pre-pilot and pilot stages. This trail of decision-making served to document the evolution of each task.

[4] This replicates Figure 4 of the report.

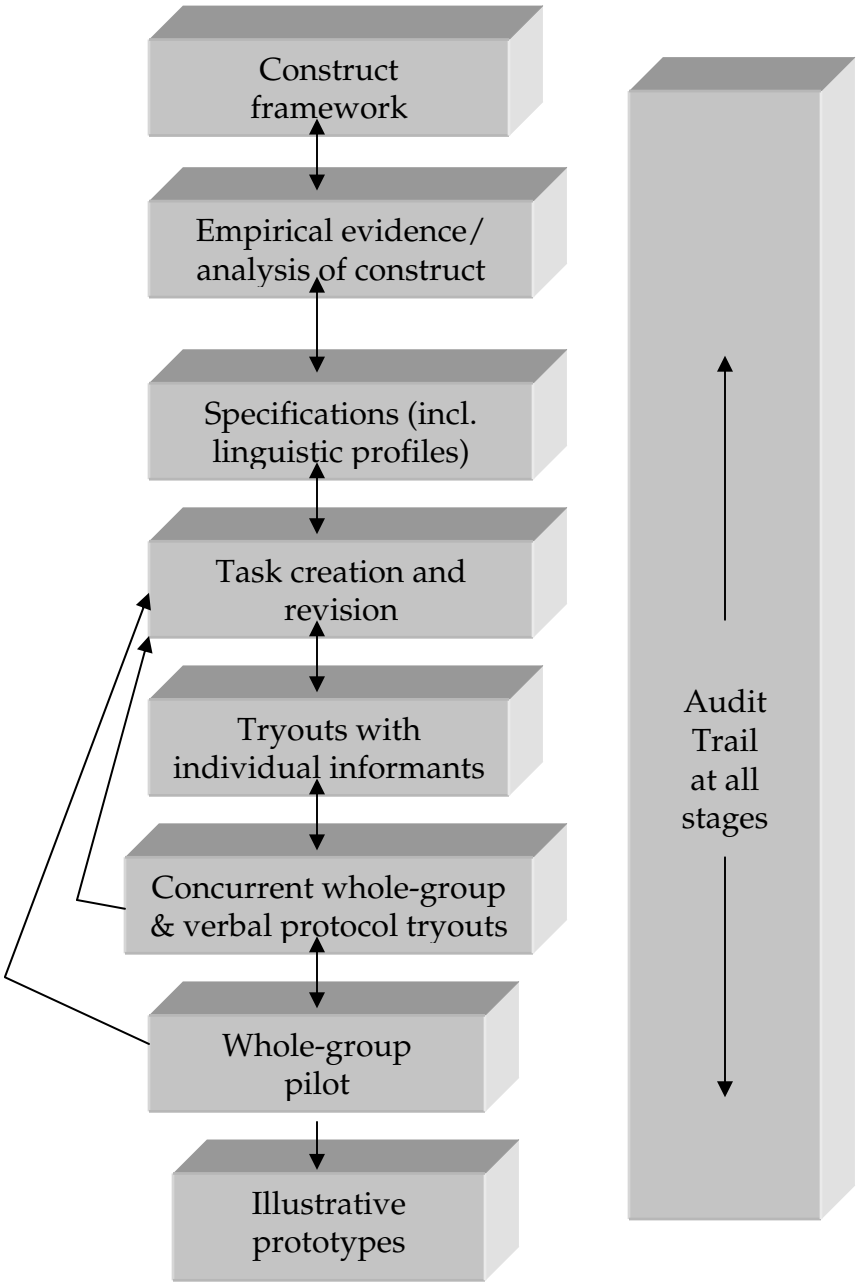


Figure 1. Proposed extensions/modifications to the test development process.

However, as Figure 1 illustrates, we recommend that the process be expanded to include an audit trail at every stage of test development, from initial construct definition all the way through to prototype creation. The return arrows show how information from the tryouts and pilot administration impacted our task revision (including rejection). However, the bidirectional arrows in the figure illustrate our suggestion that information flows back from the various phases of empirical testing

of tasks to also include the specification stage, and further back to the construct framework and its formulation so that specifications and the language construct(s) to be measured can be modified as new information comes to light as a result of tryouts and pilots (and by extension, ideally also modified based by information from the field testing of any preoperational test form).

The 40 tasks taken together are not intended to be used as a test because they were not developed as part of a specific assessment plan for a particular purpose such as redesignation. Thus they do not cover the full range of language necessary for a comprehensive evaluation of AEL in reading for such a purpose. Nevertheless, the step-by-step process described here for each task illustrates the complex and iterative nature of task development.

The work here should be viewed within the context of specific test development efforts. That is, a test being developed for a specific purpose such as redesignation or diagnosis would have a set of content requirements and specifications that operationalize the construct to be tested. The range of AEL functions and features to be assessed for a given purpose would be clearly articulated. For a grade or grade span, the appropriate functions would be identified, and then the features of the vocabulary and syntactic structure associated with those functions would be specified using academic standards and empirical evidence of classroom talk and texts.

For a test that would be part of redesignation decisions, the construct would be more broadly defined than for a classroom test in which a teacher may be focusing on one or two aspects of language. When decisions have been made about what specifically is to be tested, approaches for measuring the functions and features should be considered, specifications drafted, and tasks prepared for small-scale tryouts. The process described in this report takes potential tasks through tryout, modification, and piloting stages, and produces an audit trail for each task.

In addition to following the evolution of each task, the broad picture of the full test must be kept in mind to ensure adequate sampling of each content point being assessed. Initially, a target plan for full-test content should be prepared with the number of tasks for each function and features reflecting their importance (empirically established) within the construct. In addition, time limitations and other operational constraints should be noted. In other words, test parameters must be established. In a large-scale assessment several functions might be represented

equally, whereas in a classroom test, one or two functions may receive the most emphasis due to coverage in the curriculum during the time prior to testing. Having a full-test plan, which may be modified throughout the process, nevertheless provides a structure for guiding task development.

After small-scale tryouts, revisions, and piloting as described in the current report have been completed with a sufficient number of tasks to allow adequate content coverage according to the test design, test assembly for field testing effort begins. The field test data provide further evidence about the quality of the tasks as well as whole-test information on the reliability and validity of the instrument. Field testing provides the first evidence of how well the tasks taken as a whole function to achieve the purpose of the instrument. The more thorough the early stages of test development (as we have operationalized them in the current report) are implemented, the fewer tasks need replacing at the later field testing stage in the test development process.

To conclude, the goal of the CRESST academic English research effort has been to illustrate a process that would lead to valid and reliable instruments for assessing the English language skills of ELL students K-12. We have tried to show the importance of each step in the process, and along the way have stressed the role of empirical evidence as the foundation for developing instruments of high technical quality. The process is systematic yet flexible by allowing data to continually inform the effectiveness of tasks. Its iterative nature is the key to assuring quality assessments that are revised periodically through a feedback system. Documentation at every stage helps establish the validity argument of the assessment. Only by following a rigorous development path (with of course, on-going monitoring once an assessment is operational), can we ensure that students' language skills are being accurately and fairly evaluated. We hope that the work reported here contributes to that goal.

References

Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores in Mathematics* (CSE Tech. Rep. No. 428). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standard for educational and psychological testing*. Washington, DC: Author.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda*. Washington DC: National Academy Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bailey, A. L., & Butler, F. A. (2002/2003). *An evidentiary framework for operationalizing academic English for broad application to K-12 education: A design document.* (Final Deliverable to OERI/OBEMLA Contract No. R305B960002) (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., & Butler, F. A. (2006). A conceptual framework of academic English language for broad application to education. In A.L. Bailey, (Ed.). *Language Demands of School: Putting academic English to the test.* New Haven CT: Yale University Press.

Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2001/2004). *Towards the characterization of academic English in upper elementary science classrooms*. (Final Deliverable to OERI Contract No. R305B960002) (CSE Tech. Rep. No. 621). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., & Heritage, M., (Forthcoming). *Formative Assessment for Literacy Learning: Developing reading and academic English proficiency together, K-6*. Thousand Oaks, CA: Sage-Corwin Press.

Bailey, A. L., Stevens, R., Butler, F. A., Huang, B., & Miyoshi, J. N. (2005). *Using Standards and Empirical Evidence to Develop Academic English Proficiency Test Tasks in Reading* (CSE Tech. Rep. No. 664). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in Fifth-Grade Mathematics, Science and Social Studies Textbooks* (CSE Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., & Castellon-Wellington, M. (2000/2005). *Students' concurrent performance on tests of English language proficiency and academic achievement* (Final Deliverable to OERI, Contract No. R30B60002) (In CSE Tech. Rep. No. 663). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2003/2004). *An approach to operationalizing academic English for language test development purposes: Evidence from fifth-grade science and math* (CSE Tech. Rep. No. 626). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Butler, F. A. & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing, 18* (4), 409-427.

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic English learning approach*. Reading, MA: Addison-Wesley Publishing Company.

Cohen, A. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment In ESL.* Mahwah NJ: Lawrence Erlbaum Associates.

Davidson, F., Kim, J. T., Lee, H.J., Li, J., & Lopez, A. (2006). Using and auditing test specifications in language test development. In A. L. Bailey (Ed). *The Language Demands of School: Putting academic English to the test.* New Haven CT: Yale University Press.

Davidson, F., & Lynch, B. K. (2002). Testcraft: *A teacher's guide to writing and using language test specifications*. Newhaven, CT: Yale University Press.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and practice, 23,* 17-27.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, *64,* 55-75.

No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2002)

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Scarcella, R. (2003). *Academic English: A conceptual framework* (Tech. Rep. No. 2003-1). Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Schleppegrell, M. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431-459.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic English and content assessment: Measuring the progress of ELLs* (CSE Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

United States Government Accountability Office (2006). No Child Left Behind Act: *Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency* (GAO Publication No. GAO-06-815). Washington, DC.

Appendix

Profiles of Effective Tasks

Task Profile I - Social Studies-based

Original Draft Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

Fill in the blanks using vocabulary words from the passage.

The Hessian troops were _____(attacked)___ by George Washington.

Task Specifications

Framework category: Vocabulary

- **General description and text type:** *Students will complete a sentence using vocabulary words that are defined in a multi-paragraph expository text.*
- **Task format:** *Sentence completion using words from the passage.*
- **Stimulus attributes:** *A multi-paragraph expository text generally consisting of 3-5 paragraphs.*
- **Response attributes:** *The stimulus is followed by incomplete sentences. Students complete each sentence by filling in the blank with the correct verb from the passage.*
- **Standard addressed:** *ELD Standard addressed: Advanced Vocabulary and Concept Development; California Content Standard addressed: Social Studies: 5.5 (4)*
- **Target Academic Language Constructs:** *Specialized and general academic vocabularies are the focal linguistic features. Also measured are academic language functions: "explanation," "description", "provide instruction/guidance" and "reference to text/visual"; simple and complex grammar.*

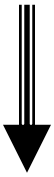
Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence (range)	10	8
Sum of Words	10	8
Total # of words (token)[a]	10	8
Total # of words (type)[b]	9	8
Lexical Features		
Academic vocabulary - general (token)	3	
Academic vocabulary - general (type)	3	
Academic vocabulary – specialized (token)		3
Academic vocabulary – specialized (type)		3
Low-frequency words (token)	1	2
Low-frequency words (type)	1	2
3-or-more-syllable words (token)	1	1
3-or-more-syllable words (type)	1	1
Avg. % of nominalizations per selection	1	
Sentence Type		
Simple sentences		1
Complex sentences	1	
Grammatical Features		
Noun phrases	3	2
Participial modifiers	1	
Passive voice verb forms		1
Prepositional phrases	1	1
Organizational Features		
Description		1
Explanation	1	
Provide instruction or guidance	1	
Reference to text or visual	1	

a. "Token" refers to the total number of words.
b. "Type" refers to the number of different words.

PHASE 1: Pre-Pilot Tryouts
Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on the feedback from phase I tryout, we decided to italicize and bold the phrase "vocabulary words from the passage" in the instructions to make it clear that ***only*** words from the passage are acceptable answers. (See highlighted area in draft task below)



Task Modified; Passage Intact for Phase 1 Pre-pilot

Modified Task

Passage

George Washington was born in 1732 in Westmoreland County, Virginia. Although his parents were landowners, they were not one of Virginia's wealthiest families. Washington was good at mathematics, but never went to college.

Washington's first job, at the age of 16, was as a surveyor. A surveyor is a person who measures land. In the middle of the 1700s many colonists were moving west and needed his services. His work paid well, and he was able to use his money to buy land.

[paragraphs omitted]

Certain of future victories, General Howe decided to rest for the winter in New York City. Washington knew that the British would not try to advance again until the spring. So he planned a surprise attack on the close to 1,400 Hessian troops in Trenton, New Jersey. The password Washington gave his soldiers was "Victory or Death!" After nightfall on Christmas Day, December 25, 1776, Washington and his troops crossed the Delaware River into New Jersey. The next morning, they surprised the Hessians, who quickly surrendered. "This is a glorious day for our country," said Washington.

Fill in the blanks using ***vocabulary words*** from the passage.

The Hessian troops were _____(attacked)_____ by George Washington.

Statistical Results from Whole Group Tryout

Item Difficulties (% correct), p = .40, (95% CI = .26-.54)
Item discrimination, D = .456

	n (Total = 45)	Percent correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 7 5^{th} = 16 6^{th} = 22	4^{th} = 29% (2) 5^{th} = 19% (3) 6^{th} = 64% (14)	Unclear	S
Gender	Girls = 24 Boys = 21	Girls = 38% (9) Boys = 48% (10)	Boys higher	NS
Home language	English = 38 Non-English = 7	English = 42% (16) Non-English = 43% (3)	Similar	NS

[a] In fact, all but 3 students in the Non-English group had Spanish as a home language (one child each for Arabic/English, Korean and Mandarin).

Breakdown of Whole Group Responses

- Correct Answer
  surprised OR attacked: 42.2% (*n*=19)
- Incorrect but Meaningful (ICM): 26.7% (*n*=12)
  - defeated (A 6^{th} grader, Korean as Home Language)
  - killed (A 5^{th} grader, Native-English speaker)
- Incorrect and Irrelevant (ICI): 31.1% (*n*=14)
  winners (A 5^{th} grader, Native-English speaker)
  ordered (A 6^{th} grader, Native-English speaker)

Verbal Protocol Analysis from Tryout

Among the 10 students who answered the item, only half provided spontaneous comments. One 4^{th} grader explained that "*because George Washington planned a surprise attack*", the answer should be "*attacked.*" Another 5^{th} grader found the answer after having gone through the passage a few times and tried different answers (*Were threatened? Surrendered maybe? Surrendered? No, "surprised".*) Generally speaking, although it did take students some time to look for the answer in the passage, those who persisted were able to arrive at the correct answer.

Excerpt from Audit Trail

This item had medium level of difficulty (p = .40) and a very good discrimination index (D = .456). Based on results from whole group tryouts, we revised our scoring rubric and included "*surprised*" as an acceptable answer. Given its reasonable difficulty level and promising discrimination index, we decided to retain this item for Phase 2 pilot.

Passage & Item Intact for Phase 2 Pilot

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct), p = .23, (95% CI = .15-.31)
Item discrimination, D = .43

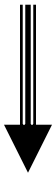
	n (Total = 111)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 15	4^{th} = 0%	Unclear	S
	5^{th} = 66	5^{th} = 34.8%		
	6^{th} = 30	6^{th} = 10%		
Gender	Girls = 59	Girls = 27.1%	Girls higher	NS
	Boys = 52	Boys = 19.2%		
Home language	English = 51	English = 33.3% (17)	English group higher than Spanish group	S
	Spanish = 60	Spanish = 10.6% (5)		
	Other = 13	Other = 30.8% (4)		

Breakdown of Pilot Group Responses

- Correct Answer
  - surprised OR attacked: 23.4% (*n*=26)
- Incorrect but Meaningful (ICM): 2.7% (*n*=3)
  - defeated (A 5^{th} grader, Spanish Home Language)
  - killed (A 6^{th} grader, Spanish Home Language)
- Incorrect and Irrelevant (ICI): 73.9% (*n*=82)
  - winners (A 6^{th} grader, Native-English speaker)
  - against (A 5^{th} grader, Native-English speaker)
  - surrendered (A 4^{th} grader, , Spanish Home Language)

Excerpt from Audit Trail

Compared to the results from Phase I pre-pilot studies, this item yielded item difficulty index in Phase 2 Pilot (p = .23) that suggested it was more difficult for students. However, the discrimination index remained very promising (D = .43), indicating that the item distinguished correctly between good and poor readers. Also, the item significantly distinguished between the performance of students with different language backgrounds. Students from English language backgrounds performed significantly better than those who had Spanish as a home language.



VERDICT: ***Passage & Task Retained as an AELP Prototype***

Summary of Effective Task Profile I

The social studies-based task targeted at students' comprehension of AEL vocabulary and grammar through a sentence completion task. Students first read a multi-paragraph expository text taken from a social studies textbook, and were requested to then identify a verb from the passage to fill in the blank of an incomplete sentence. This item addresses both ELD standard (Advanced Vocabulary and Concept Development) and California Content Standard for Social Studies 5.5 (4).

The linguistic analysis shows that the constructs of both the task stem/prompt and task response pertained predominantly to knowledge of specialized and general academic vocabulary. The task stem also tapped knowledge of simple and complex English grammar, and academic language functions *explanation, provide instruction or guidance*, and *reference to text*. The task response also required knowledge of AEL function *description*.

The student informant suggested that we italicize and bold the phrase "vocabulary words from the passage" in the instructions to make it clear that ***only*** words from the passage are acceptable in the responses. The passage remained intact and the instructions were modified accordingly for the pre-pilot. Statistical results from these tryouts revealed that the task was of medium difficulty level (p = .40), and had a very good discrimination index (D = .456).

We decided to retain this item for the Phase 2 pilot because of its reasonable difficulty level and promising discrimination index. Possibly due to sample background differences, students in the Phase 2 pilot did not perform as well as their counter-parts in the Phase 1 tryouts; item difficulty increased (p = .23). However, the task still discriminated effectively between good and poor readers, as well as distinguished between students with English and Spanish home language backgrounds.

Task Profile II - Math-based

Original Draft Task

Passage

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies?

What is the word problem asking about?

a) How much Carlotta spent on supplies.
b) How many packages of lemonade she sold.
c) How much profit Carlotta made.*
d) How much lemonade costs.

*correct response

Task Specifications

Framework category: Demonstration of Comprehension (through paraphrase)

- **General description and text type:** *Students will identify the problem statement in a mathematics word problem and select the correct paraphrase from multiple-choice sentence options*
- **Task format:** *'Wh' question with multiple-choice sentence options*
- **Stimulus attributes:** *A mathematics word problem generally of 2-3 sentences in length with a problem question or imperative statement at the end. (empirical evidence) The target academic language function construct is "paraphrase", which requires the processing of the same idea expressed in different words.*
- **Response attributes:** *Circle the correct multiple-choice option from the four options provided.*
- **Standard addressed:** *ELD Standard addressed: Early Advanced Comprehension and Analysis; California Content Standard addressed: Math Number Sense 2.0 (2.1)*
- **Target Academic Language Constructs:** *Academic language functions "paraphrase" and "summarize"; and specialized academic vocabulary*

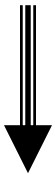
Linguistic Analysis Profile

	Stem/Prompt	Response
Descriptive analysis		
(Mean) no. of words per sentence(range)	7	5.5 (4-7)
Sum of Words	7	22
Total # of words (token)	7	22
Total # of words (type)	7	15
Lexical features		
Academic vocabulary - specialized(token)	2	2
Academic vocabulary - specialized(type)	2	2
3-or-more-syllable words(token)		3
3-or-more-syllable words(type)		2
Derived words (token)		2
Derived words (type)		2
Sentence type		
Simple sentences	1	NA
Other sentence types		4 clauses
Grammatical features		
Prepositional phrases		1
Organizational features		
Paraphrase	1	1
Question	1	
Summary	1	1

PHASE 1: Pre-Pilot Tryouts

Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on feedback from the tryout, a note of caution was added to the end of the reading passage to prevent students from working on the math problem.



Passage Modified; Item Intact for Phase I Pre-pilot

Modified Task

Passage

Carlotta bought 9 packages of lemonade for $1.10 each and 2 packages of cups for $1.09 each. She sold 23 cups of lemonade every hour for 4 hours at $0.40 per cup. How much more money did Carlotta earn than she spent on supplies? [DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.][a]

What is the word problem asking about?

a) How much Carlotta spent on supplies.
b) How many packages of lemonade she sold.
c) How much profit Carlotta made.*
d) How much lemonade costs.

[a] The highlights of modifications only appear in the example passage and draft task presented above for demonstration purpose. They were removed from the version students receive in pre-pilot and pilot testing.

Statistical Results from Whole Group Tryout

Item Difficulties (% correct) p = .68 (95% CI = .56-.80)
Item discrimination (D) = .332

	n (Total = 75)	Percent Correct (Raw Number)	Trend	Statistical Significance
Grade	4th = 21 5th =27 6th = 27	4th = 54% 5th = 74% 6th = 78%	Positive	NS
Gender	Girls = 36 Boys = 39	Girls = 78% Boys = 62%	Girls higher	NS
Home Language	English = 61 Non-English = 14	English = 74% (45) Non-English = 50% (7)	English group higher	NS

Breakdown of Tryout Group Responses

- 69.3% (*n*=52) chose the correct answer C: How much profit Carlotta made.
- 13.3% (*n*=10) chose distractor answer A: How much Carlotta spent on supplies.
- 5.3% (*n*=4) chose distractor answer B: How many packages of lemonade she sold.
- 1,3% (*n*=1) chose distractor answer D: How much lemonade costs
- 10.6 % (*n*=8) given opportunity to work on the item but provided no response

Verbal Protocol Analysis from Tryout

Among the students who had provided comments on this item (*n*=13), some of them had specifically identified either the name "Carlotta" in the word problem or the word "profit" in the correct answer as difficult words. Although theoretically unfamiliarity with the proper noun "Carlotta" would not impede reading comprehension, some students would pause at the word and make extra efforts to pronounce the word correctly. About half of the students who had answered the question got the correct answer (*n*=6 out of 13). The strategies those students reported included going back to the passage (I really didn't exactly understand so I went back up to the passage and read the question that they asked, so then I noticed that profit is basically the same thing earned of...how much she's earned…so it means how much profit…so profit means the same thing. Comment from a 6th grader) and eliminating answers (I just eliminated…[unintelligible]. How many packages of lemonade she sold, it doesn't say that there, how much Carlotta spent on supplies, and the problem doesn't really say that. It said how much more money did Carlotta earn than she spent on supplies. So that's different. Comment from a 4th grader). On the other hand, based on the comments from students who had answered the question incorrectly, it appeared that some of them still treated the question as more of a "math" word problem than a reading comprehension item of their academic English proficiency. For example, a 4th grader chose the wrong answer "a) How much Carlotta spent on supplies" and rationalized his answer as following: *Because it tells you all the prices for sure. It's not how many packages of lemonade she sold… How much profit Carlotta made… It doesn't even tell you that.*

Excerpts from Audit Trail

Although this item had a relatively low difficulty index (p=.68), it reasonably discriminated among good and poor readers (D=.332). In addition, it also distinguished across grade and home language background. Review of the student responses revealed that the distractors were also plausible and effective.

Passage & Item Intact for Phase 2 Pilot

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct) p = .42 (95% CI = .34-.51)
Item discrimination (D) = .37

	n (Total = 125)	Percent correct (Raw Number)	Trend	Statistical Significance
Grade	4^{th} = 18 5^{th} = 76 6^{th} = 31	4^{th} = 33.37% 5^{th} = 47.4% 6^{th} = 35.5%	Unclear	NS
Gender	Girls = 67 Boys = 58	Girls = 47.8% Boys = 36.2%	Girls higher	NS
Home Language	English = 52 Spanish = 56 Other = 17	English = 55.8% (29) Spanish = 32.1% (18) Other = 35.3% (6)	English group higher than both Spanish and Other	S (English higher than Spanish)

Breakdown of Pilot Group Responses

- 42.4% (*n*=53) chose the correct answer C: How much profit Carlotta made.
- 38.4% (*n*=48) chose distractor answer A: How much Carlotta spent on supplies.
- 12% (*n*=15) chose distractor answer B: How many packages of lemonade she sold.
- 5.6% (*n*=7) chose distractor answer D: How much lemonade costs
- 1.6 % (*n*=2) given opportunity to work on the item but provided no response

Excerpts from Audit Trail

The item difficulty level changed from .68 to .42 in the Phase 2 pilot findings, suggesting it was harder for the pilot students. The item discrimination index remained adequate. The task also significantly distinguished between students with English and Spanish home language backgrounds.

VERDICT: ***Passage & Task Retained as an AELP Prototype***

Summary of Effective Task Profile II

This math-based task was created to measure students' knowledge of English grammar and discourse through a multiple-choice task. Students encountered a word problem of 2-3 sentences in length. They are then required to select the correct answer from four options that answered the main-idea question. This required students to understand that "how much more" in the passage is, in this context, equivalent to the word "profit" in the correct response. This item addressed both ELD standard (Early Advanced Comprehension and Analysis) and California Content Standard for Math (Number Sense 2.1).

The linguistic analysis reveals that the task stem/prompt and the response involved knowledge of the *paraphrase* academic language function or organizing feature, as well as the *summary* function. Knowledge of specialized academic vocabulary is also required for both the stem/prompt and response. The student informant suggested that a note be added to the end of the passage to refrain students from working on the math problem. The item remained intact and the passage was modified accordingly for whole group tryout.

Statistical results from pre-pilot revealed that the item had a low difficulty index (p = .68). However, it reasonably discriminated among good and poor readers (D = .332) and distinguished across grade and home language background. The distractors were also shown to be plausible and effective. We thus retained the task for the Phase 2 pilot. Similar to the findings of the first effective AELP task above, students in the Phase 2 pilot (p = .42) performed less well than those in the Phase 1 tryouts. However, the task maintained an adequate discrimination index (D= .37). It also significantly distinguished between students with English and Spanish home language backgrounds.

Task Profile III - Math-based

Original Draft Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster?

Read the problem. Then complete the table.

Person	Year	Distance	Days	From	To
Stilt walker #1	1980		158		Bowen, Kentucky
Stilt walker #2	1891	1830 miles		Paris, France	***Target Item (Moscow, Russia)***

Task Specifications

Framework category: Academic Language Function "Comparison", and Vocabulary

- **General description and text type:** *Students will read the word problem and retrieve appropriate information from the text to fill in the gaps in a table.*
- **Task format:** *Fill in the gaps in a table using information from the text.*
- **Stimulus attributes:** *A mathematics word problem generally of 2-3 sentences in length with a problem question or imperative statement at the end.*
- **Response attributes:** *Fill in the blanks in a table by retrieving requested information from the text.*
- **Standard addressed:** ELD Standard addressed: *Advanced Reading Comprehension*; California Content Standard addressed: *Math Number Sense 2.0 (2.3).*
- **Target Academic Language Constructs:** *Focal academic language function: "comparison," also "scenario," "labeling", "provide instruction/guidance" and "summarize;" and general academic vocabulary.*

Linguistic Analysis Profile

	Stem/Prompt	*Response*
Descriptive Analysis		
(Mean) no. of words per sentence(range)	3.5 (3-4)	NA
Sum of Words	7	21
Total # of words (token)[a]	28	21
Total # of words (type)	25	17
Lexical Features		
Academic vocabulary - general (token)	4	
Academic vocabulary - general (type)	4	
Low-frequency words (token)	7	1
Low-frequency words (type)	5	1
3-or-more-syllable words(token)		1
3-or-more-syllable words(type)		1
Derived words (token)		3
Derived words (type)		2
Avg. % of nominalizations per selection		1
Sentence Type		
Simple sentences	2	
Grammatical Features		
Noun phrases	3	16
Organizational Features		
Scenario		1
Comparison		1
Labeling		1
Provide instruction or guidance	1	
Reference to text or visual	1	1

[a]This frequency also includes column headings and content of the table in the task.

PHASE 1: Pre-Pilot Tryouts

Initial Feedback on Task Formatting and Directions

The student informant answered the task correctly. Based on the feedback from the student and notes from our internal review meeting, we had made a few changes to the format of the passage and the item. We added a cautionary note at the end of the math word problem to prevent students from treating the item as a mathematical question. Instead of treating the whole table as one item, we separated each blank in the table into individual items. Additionally, we modified the instruction from "Read the problem. Then complete the table" to "Fill in the blanks for questions 1 through 4 in the table below," and accordingly added lines and corresponding numbers for each blank in the table for students to fill in their responses.

Passage & Item Modified for Phase 1 Pre-pilot

Modified Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? [DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]

Fill in the blanks for questions 1 through 4 in the table below.

Person	Year	Distance	Days	From	To
Stilt walker #1	1980	(1.) _____	158	(2.) _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	(3.) _____	Paris, France	(4.) (Moscow, Russia) _____

Statistical Results from Whole Group Tryout

Item Difficulties (% correct) p = .90 (95% CI = .82-.98)
Item discrimination (D) = .263

	n (Total = 69)	Percent correct (raw number)	Trend	Statistical significance
Grade	4^{th} = 17 5^{th} = 27 6^{th} = 25	4^{th} = 88% 5^{th} = 93% 6^{th} = 100%	Positive	NS
Gender	Girls = 37 Boys = 32	Girls = 95% Boys = 94%	Similar	NS
Home Language	English = 56 Non-English = 13	English = 93% (52) Non-English = 100% (13)	Spanish group higher	NS

Breakdown of Tryout Group Responses

- Correct Answer = Moscow, Russia: 94.2% (*n*=65)
- Incorrect but Meaningful (ICM): 2.9% (*n*=2)
  - Russia (A 4^{th} grader, both Native-English speaker)
- Incorrect = 2.9% (*n*=2)
  - Paris, France, Moscow, Russia (A 5^{th} grader, Native-English speaker)
  - Paris, France. (A 5^{th} grader student, Native-English speaker)

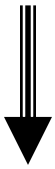
Verbal Protocol Analysis from Tryout

All of the fourteen students who worked on the question answered it correctly, and most of them also answered the question promptly. However, only half of them had provided comments on the item either spontaneously or with prompts.

Specifically, five of the comments pertained to the format of the item and familiarity with the item type (i.e. tabular form). One 6^{th} grader initially had difficulty understanding the task, but was able to answer the questions after the researcher's explanation. Another 5^{th} grader commented on the limited space for writing down her answer (*I don't think there'd be enough room to write the whole entire name...[referring to place name]*). On the other hand, two students reported using the strategy of going back to the passage to look for the answer (*I first read all the stuff that they said and then I went back to the passage to look what the answers were*. Comment from a 6^{th} grader).

Excerpts from Audit Trail

Although this item turned out to be quite easy for the students (p = .90), it nonetheless had a fair item discrimination index (D = .263). It also distinguished across grades and yielded similar performances for girls and boys. We thus decided to retain this item for the Phase 2 pilot with some modifications to the item. The modifications included adding a title to the table to make the task more transparent and italicizing and reformatting each question number in the table.



Item Modified; Passage Intact for Phase 2 Pilot

Modified Task

Passage

In 1980, a man walked 3,008 mi on stilts from Los Angeles to Bowen, Kentucky. The trip took 158 days. In 1891, a stilt walker traveled from Paris, France, to Moscow, Russia, going 1,830 mi in about 54 days. Who traveled faster? **[DO NOT ANSWER THIS WORD PROBLEM. INSTEAD, ANSWER THE QUESTIONS BELOW ABOUT THIS PASSAGE.]**

Fill in the blanks for questions *(1)* through *(4)* in the table below.

Table: Stilt Walker Travel by Distance and Days

Person	Year	Distance	Days	From	To
Stilt walker #1	1980	*(1)* _____	158	*(2)* _____	Bowen, Kentucky
Stilt walker #2	1891	1,830 miles	*(3)* _____	Paris, France	*(4)* _____

Statistical Results from Phase 2 Pilot

Item Difficulties (% correct) p = .75 (95% CI = .67-.83)
Item discrimination (D) = .43

	n (Total =111)	Percent correct (raw number)	Trend	Statistical significance
Grade	4^{th} = 15 5^{th} = 66 6^{th} = 30	4^{th} = 53% 5^{th} = 80% 6^{th} = 73%	Unclear	NS
Gender	Girls = 59 Boys = 52	Girls = 81% Boys = 67%	Girls higher	NS
Home Language	English = 51 Spanish = 47 Other = 13	English = 88% (45) Spanish = 64% (30) Other = 62% (8)	English group higher than Spanish and other group	S

Breakdown of Pilot Group Responses

- Correct Answer = Moscow, Russia: 74.8% (*n*=83)
- Incorrect but Meaningful (ICM): 1% (*n*=1)
  - Russia (A 6th grader, Spanish Home Language)
- Incorrect = 24.3% (*n*=27)
  Paris, France (A 5th grader, Spanish Home Language)
  Los Angeles (A 6th grader, Spanish Home Language)
  1830 (A 5th grader, Spanish Home Language)

Excerpts from Audit Trail

The item difficulty level changed from .90 in the pre-pilot phase to .75 in the pilot phase, suggesting fewer students answered correctly. The finding was likely due to the differences in sample demographics: there were more ELL students in the pilot. However, the item had a higher discrimination index (D = .43) and distinguished between students from different home language backgrounds.

VERDICT: Passage & Task Retained as an AELP Prototype

Summary of Effective Task Profile III

This math-based item was intended to measure students' knowledge of discourse and vocabulary through a graphic organizer task in a table format. Students first read a word problem of 2-3 sentences in length. They were then required to retrieve information from the word problem to fill in the blanks in a table. This item addressed both ELD standard (Early Advanced Reading Comprehension) and California Content Standard for Math (Number Sense 2.3).

The linguistic analysis reveals that the task stem/prompt construct tapped into knowledge of academic language functions *provide instruction or guidance* and *reference to text*. The task response pertained not only to these functions, but a focal function *comparison*, and additional functions *labeling* and *scenario*, as well as general academic vocabulary.

A few changes in format were made to the task for the pre-pilot based on student informant feedback and internal review meeting notes. Statistical results from Phase 1 tryouts revealed that the task was easy (p = .90). However, it moderately discriminated among good and poor readers (D = .263) and distinguished across grades. The task was further modified as described above and was retained for the Phase 2 pilot. Comparing findings from the Phase 1 tryouts and the Phase 2 pilot, the item difficulty level changed from .90 to .75, suggesting it was more difficult for the pilot sample possibly due to differences in sample demographics. However, the task adequately discriminated between good and poor readers (D = .43), as well as distinguished between students from different home language backgrounds.